The Project ARIANE: Conceptual Queries to Information Databases

Michel Joubert, Ph.D., Jean-Jacques Robert, Ph.D., François Miton, M.D., M.S.,
Marius Fieschi, M.D., Ph.D.
CERTIM. Faculté de Médecine. Boulevard Pierre Dramard.
F-13326 Marseille Cedex 15. France

As information databases we consider all the collections of data records indexed by key-words, stored and delivered by computer systems. In previous research works we demonstrated the interest to design a conceptual model, in the conceptual graphs formalism, and to implement a computational model for information retrieval in large information databases. These models are based on the UMLS knowledge sources. This paper reminds briefly these models and describes tests done in querying a patients database and a bibliographical database.

INTRODUCTION

As information databases we consider all the collections of data that are stored and delivered by computer systems, such as patients databases, bibliographical references databases, documents servers, and so on. We will suppose that the entities the users want to retrieve are indexed by key-words chosen in a controlled vocabulary. Information retrieval in large databases is a process which consists in a series of queries and refinements. This non-deterministic search is due to the flat organization of information, even if it is indexed by tree-structured thesauri. The main problem encountered by end users in querying information databases is to map their own perception of concepts to their representation in computer systems.

Let's suppose a physician is interested in the research works about the efficiency of AZT in AIDS prevention[1]. A traditional query could be: "AIDS AND AZT". In the semantic network of UMLS[2], AIDS is a "Disease or Syndrome" and AZT is a "Pharmacological Substance". Semantic relationships between these types of concepts are: "treats", "prevents", "causes" and "complicates". The selection of "prevents" may produce the query formula: (*AIDS/ prevention & control) AND (*AZT/ therapeutic use), according to the MeSH syntax. This simple example demonstrates how a knowledge of the domain makes the dialog easier between a end-user and a computer system. The use of the semantic relationship between the two concepts has specified the context in which each concept must be considered. Instead of a query formula involving key-words only, without semantics, we propose to express a query by a data structure that involves concepts and relationships. This structure is a deep representation of a natural language sentence.

Even if the conceptual graphs theory has been initially designed for natural language semantical representation and processing[3-6], it is powerful enough to represent knowledge efficiently. It has been used to build classification systems of medical concepts[7] and to represent clinical data[8-10]. In previous works we proposed a conceptual model able to assist users to build queries to information databases, and a computational model to implement this model[11-12]. The first section summarizes our approach. The following section describes the tests we have done with information databases.

CONCEPTUAL AND COMPUTATIONAL MODELS OF UMLS COMPONENTS

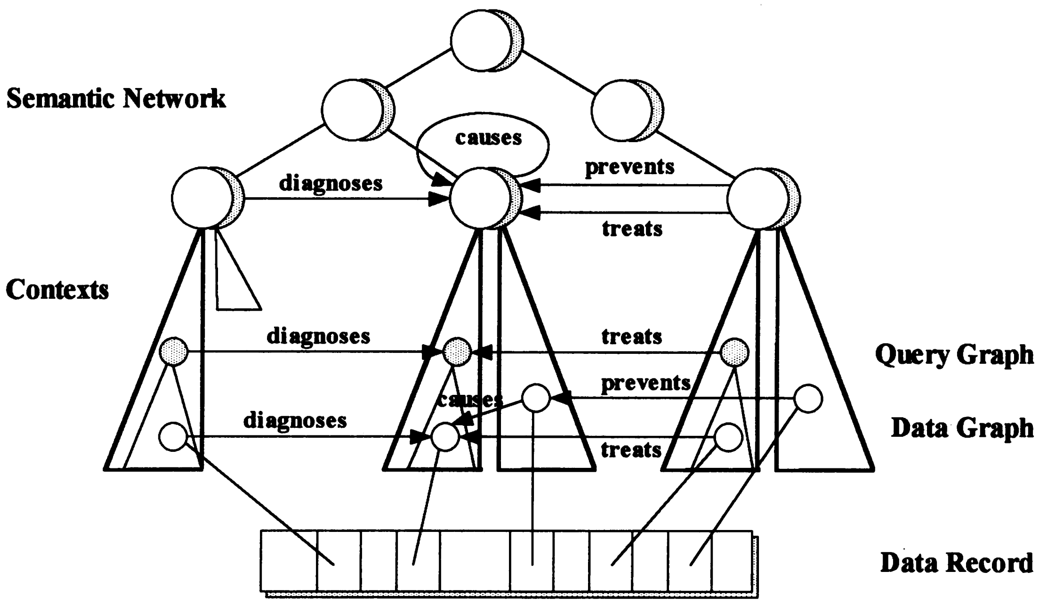
The Unified Medical Language System

The Unified Medical Language System (UMLS) is a wide project of the U.S. National Library of Medicine (NLM)[13]. The UMLS is a complex collection of medical concepts, terms and relationships issued from standard classifications. Three main components constitute the UMLS data structure: the so-called Metathesaurus (Meta-1), the Semantic Network, and the Information Sources Map. The core concepts which have been isolated in Meta-1 are connected to generic types of concepts in the Semantic Network. The types of concepts may be interconnected by binary semantic relationships.

The data structure of Meta-1 is based on hierarchies and associations. The association relationship links a term to related terms and to a preferred term. The hierarchies structure the preferred terms into more generic terms and more specific ones. This hierarchical relationship divides Meta-1 into several micro-thesauri, according to a local specificity of concepts. These micro-thesauri represent the various viewpoints from which medical concepts can be considered.

0195—4210/96/$5.00 © 1996 AMIA, Inc.

Figure 1: building a conceptual query from semantic knowledge and contexts, and matching it with data



A Conceptual Model of UMLS Components

A first objective of the conceptual model is to represent the complex data structure of the UMLS components. We organized the components according to a data structure conceived for semantic networks exploitation. When applied to the UMLS components, this structure shows the three following levels:

- the core concepts identified in Meta-1 constitute the lowest level,
- instances of concepts are organized in contexts at an intermediate level, these contexts are the micro-thesauri of Meta-1,
- the upper level is the ontology of types of concepts of the UMLS Semantic Network.

A given concept may have various instances in different contexts, but only one in a given context. Each context is connected to one, and only one, type of concepts at the upper level. The intermediate and upper levels are represented in Figure 1.

We designed a dictionary of concepts to register information related to contexts[14]. The aim of such a dictionary is to describe the core concepts and their instances in various contexts. It appears from the representation we done that a given instance of a concept is characterized by the path followed to reach it. This means that an instance of a concept is characterized by its semantic environment: fathers, siblings and children in contexts. A unit of the dictionary contains the definition of all the instances of each core concept in its possible contexts. Such a representation guarantees that one, and only one, occurrence of each core concept in stored in the dictionary and that it encapsulates the description of the links to its possible instances.

A second objective of the conceptual model is to provide users with the capability to build associations, as conceptual graphs, which involve instances of concepts in contexts interconnected by semantic relationships inherited from the semantic network. Operations on conceptual graphs provide the users with the capability to build complex graphs which translate full sentences. The diagram of Figure 1 illustrates how semantic relationships of the upper level (Semantic Network) are used to connect instances of concepts at the intermediate level (Contexts) in order to create a query graph. The "join" operation is used to combine graphs which contain concepts in a same lineage.

The matching process between a query graph and records in an information database is also processed by conceptual graphs operations. Since data records are indexed, it is possible to constitute the list of the instances of concepts in relation to the content of the records. To build the data graphs that are to be compared with a query graph, three cases may arise according to how the database is built:

- relationships between instances of concepts are expressed in the records, in this case nothing has to be done since each record contains its data graph,
- relationships between instances of concepts are not expressed in the records, but it is possible to

put them with certainty (e.g., an angiography and a coronary arteriosclerosis must be linked by "diagnoses"),

- relationships between instances of concepts are not expressed in the records, and it is not possible to put them with certainty.

In the first two cases, the "projection" of the query graph into the data graph of a record operates the matching. In the last case, only the instances of concepts are compared, without using the data structure given by a query graph. In each case, the structure of the contexts is used to verify if an element of a data graph is more specific than (or at least equal to) an element of the query graph. This process is illustrated by the Figure 1. We describe below tests we made in the first and third cases. The second case has been treated in a previous work [15].

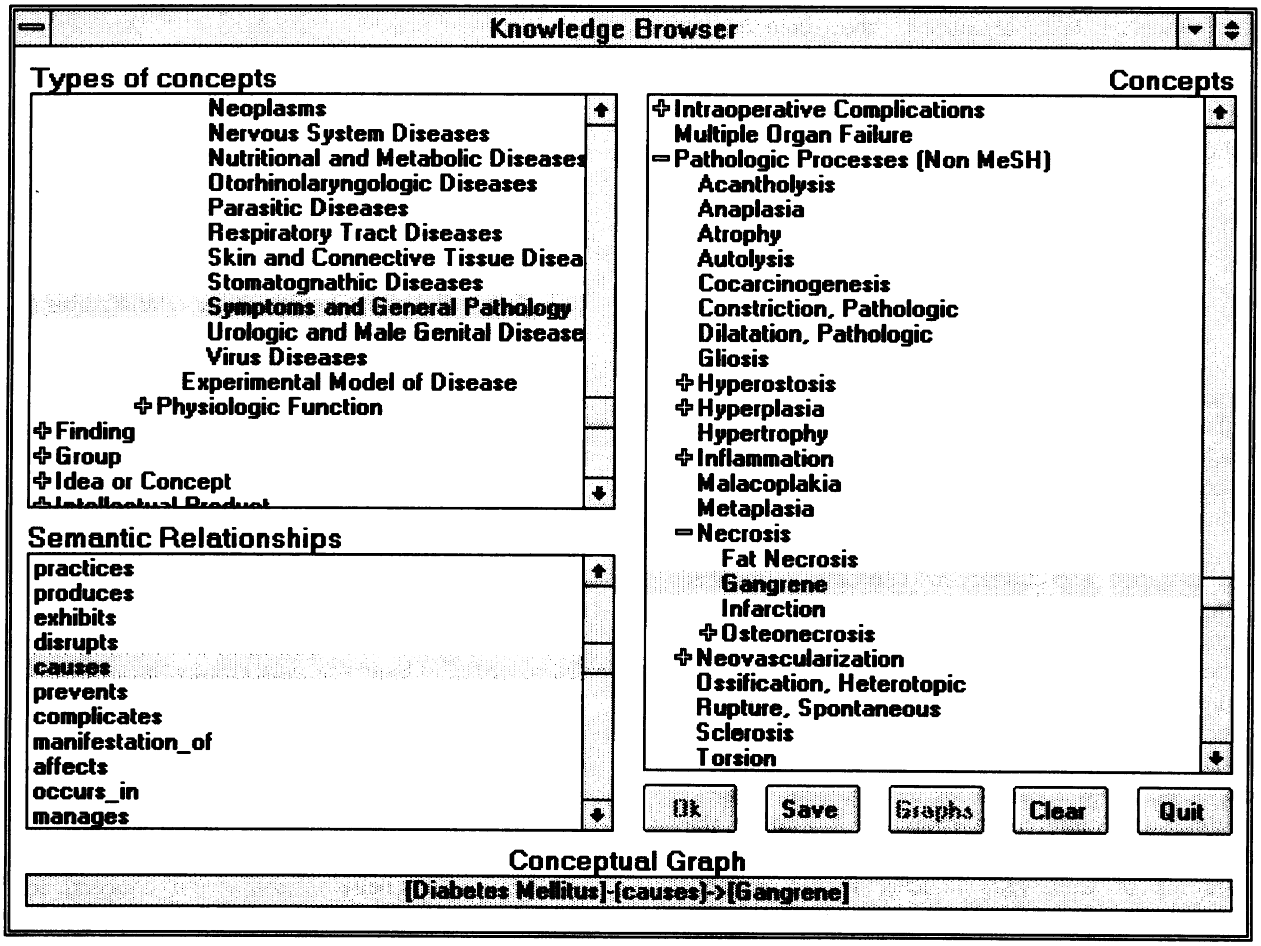
A Computational Model of UMLS Components

A typical exploitation of the above conceptual model intended to build a query is as follows:

- select a concept,
- select a context for this concept: the related type is then selected automatically,
- select a relationship involving this type: the destination type is then selected,
- select a context connected to this last type,
- select an instance of a concept in this context.

This sequence of operations builds an elementary conceptual graph which links two nodes, the two selected instances of concepts, by the means of a selected semantic relationship. Iterations of such a sequence allow to build complex connected graphs which involve instances of concepts linked by various relationships. Figure 2 shows a screen dump of the knowledge browser we have implemented. In this instance, the elementary graph "Diabetes Mellitus causes Gangrene" is saved. Its join with the graph "Necrosis co-occurring with Arterial Occlusive Disease" produces the query graph "Gangrene caused by Diabetes Mellitus and co-occurring with Arterial Occlusive Disease". This operation is possible since a gangrene is a kind of necrosis.

Figure 2: building conceptual queries from semantic knowledge and contexts



Types of concepts

Neoplasms
Nervous System Diseases
Nutritional and Metabolic Diseases
Otorhinolaryngologic Diseases
Parasitic Diseases
Respiratory Tract Diseases
Skin and Connective Tissue Disea
Stomatognathic Diseases
Symptoms and General Pathology
Urologic and Male Genital Disease
Virus Diseases
Experimental Model of Disease
+ Physiologic Function
+ Finding
+ Group
+ Idea or Concept
+ Intellectual Product

Semantic Relationships

practices
produces
exhibits
disrupts
causes
prevents
complicates
manifestation_of
affects
occurs_in
manages

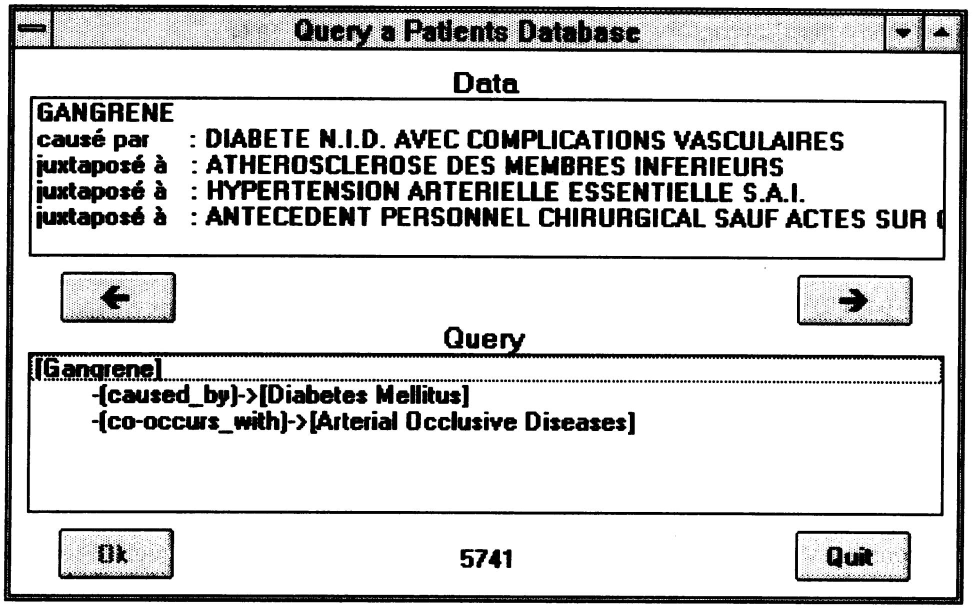
Concepts

+ Intraoperative Complications
Multiple Organ Failure
− Pathologic Processes (Non MeSH)
Acantholysis
Anaplasia
Atrophy
Autolysis
Cocarcinogenesis
Constriction, Pathologic
Dilatation, Pathologic
Gliosis
+ Hyperostosis
+ Hyperplasia
Hypertrophy
+ Inflammation
Malacoplakia
Metaplasia
− Necrosis
Fat Necrosis
Gangrene
Infarction
+ Osteonecrosis
+ Neovascularization
Ossification, Heterotopic
Rupture, Spontaneous
Sclerosis
Torsion

Conceptual Graph

[Diabetes Mellitus]-(causes)->[Gangrene]

Figure 3: applying a conceptual query to a patients database



GANGRENE
causé par : DIABETE N.I.D. AVEC COMPLICATIONS VASCULAIRES
juxtaposé à : ATHEROSCLEROSE DES MEMBRES INFERIEURS
juxtaposé à : HYPERTENSION ARTERIELLE ESSENTIELLE S.A.I.
juxtaposé à : ANTECEDENT PERSONNEL CHIRURGICAL SAUF ACTES SUR

```
[Gangrene]
     -[caused_by]->[Diabetes Mellitus]
     -[co-occurs_with]->[Arterial Occlusive Diseases]
```

The knowledge base issued from UMLS has been imported into a relational database management system (Microsoft Access). The patients database and bibliographical database we have tested the system with have been imported into the same database management system. We developed software modules with both Microsoft Visual C++ (the full version) and Microsoft Visual Basic (a "lite" version).

QUERIES TO INFORMATION DATABASES

Queries to a Patients Database

We tested the query process with a database of patients who were admitted in a general surgery ward during a four years period and for whom valid standardized discharge records have been stored. The relational table contained more than 7,000 rows. The fields under consideration are the main discharge diagnosis and associated diagnoses. The database has been constituted by surgeons who linked the associated diagnoses to the main diagnosis by the means of four relationships: "co-occurs with", "causes", "complicates", and "associated with", as they are defined in the UMLS.

Figure 3 shows a data record which satisfies the previous query: "Gangrene caused by Diabetes and co-occurring with Arterial Occlusive Disease". The main diagnosis in this data record is gangrene. The relationship "caused by" (in French: "causé par") links the gangrene to a non-insulin-dependent diabetes. The relationship "co-occurs with" (in French: "juxtaposé à") links the gangrene with an atherosclerosis located in the legs. Two other diagnoses are linked to the gangrene by the same relationship. These diagnoses and relationships compose a conceptual data graph in which the query graph is projected successfully since a non-insulin-dependent diabetes is a kind of diabetes mellitus, and an atherosclerosis is a kind of arterial occlusive disease. Moreover, the matching process between to graphs does not have to take into account the ranks of the data in the relational database, such as a traditional SQL query does.

Queries to a Bibliographical Database

A second test has been conducted with bibliographical references. We extracted from Medline all the records related to coronary diseases referenced during a year. In this case, the UMLS knowledge cannot be used directly as a data model. Since, for instance, a reference may contain key-words for several diseases and the UMLS Semantic Network contains relationships such as "causes", "complicates" or "co-occurs with", it is not possible to build a correct data graph automatically. In such a case we proceeded differently. First, we made the assumption that a query is medically correct: relationships do not connect irrelevant concepts. Then, for each data record, all the concepts present in the query graph are compared to the key-words. The concept is matched if at least a key-word is the instance of a more specific concept. The data record is selected if all the concepts of the query graph are matched by key-words.

The previous query extracts references from the documentary database. A typical instance is the following. Title: "Very distal bypass for salvage of the severely ischemic extremity". Abstract: "Forty-six bypass grafts to tibial distal to the ankle were performed in 35 patients for salvage of extremities threatened by gangrene or ... Most patients were diabetic, with severely calcified arteries, ..." Among the key-words are: "Gangrene", "Diabetic Angiopathies", etc.

DISCUSSION

Querying large medical information databases is often problematic due to the complexity of the biomedical domain. Thus, the needs for intelligent user's interfaces to information servers seems clear. We proposed a model and an exploitation tool in order to help users to express queries to such databases. The conceptual graphs theory provides with the capability to represent both data and queries in a unique formalism. We illustrated the matching process between a query graph and data graphs in two cases. When data are structured by the means of semantic relationships, the matching process operates utterly. In such a case, the use of conceptual queries for information retrieval is of great significance. When querying unstructured databases from records of which data graphs cannot be constituted, a graph query is restricted to a list of concepts which operates as an implicit conjunction of concepts. Nevertheless, the matching of concepts processes automatically the "explosion" function that some thesaurus-based information retrieval systems offer.

Acknowledgments

This work has been partially founded by the French Ministry of Education. The authors thank the U.S. National Library of Medicine which provided them with the UMLS material, and Dr Borelli (Cannes Hospital, France) who provided them gracefully with the patients database used for tests.

References

1. Lindberg DAB, Humphreys BL. The UMLS Knowledge Sources: Tools for Building Better User Interfaces. Proc. 14th SCAMC. Miller RA editor. IEEE Computer Society Press. 1990: 121-25.
2. McCray AT. The UMLS Semantic Network. Proc. 13rd SCAMC. Kingsland LC editor. IEEE Computer Society Press. 1989: 503-7.
3. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Addison Wesley. 1984.
4. Sowa JF. Toward the Expressive Power of Natural Language. Principles of Semantic Networks: Exploration in the Representation of Knowledge. Sowa JF editor. Morgan Kaufmann. 1991: 157-89.
5. Volot F, Zweigenbaum P, Bachimont B and al. Structuration and Acquisition of Medical Knowledge Using UMLS in the Conceptual Graphs Formalism. Proc. 17th SCAMC. Safran C editor. McGraw-Hill. 1993: 710-14.
6. Baud RH, Rassinoux A-M, Scherrer J-R. Natural Language Processing and Semantical Representation of Medical Texts. Methods Inf Med. 1992; 31: 117-25.
7. Bernauer J. Subsumption Principles Underlying Medical Concept Systems and their Formal Reconstruction. Proc. 18th SCAMC. Ozbolt JG editor. JAMIA Symposium supplement. 1994: 140-44.
8. Bernauer J. Conceptual Graphs as an Operational Model for Descriptive Findings. Proc. 15th SCAMC. Clayton PD editor. McGraw-Hill. 1991: 214-18.
9. Campbell KE, Musen MA. Representation of Clinical Data Using SNOMED III and Conceptual Graphs. Proc. 16th SCAMC. Frisse ME editor. McGraw-Hill. 1992: 354-58.
10. Campbell KE, Das AK, Musen MA. A Logical Foundation for Representation of Clinical Data. JAMIA. 1994; 1: 218-32.
11. Joubert M, Fieschi M, Robert JJ. A Conceptual Model for Information Retrieval with UMLS. Proc. 17th SCAMC. Safran C editor. McGraw-Hill. 1993: 715-19.
12. Robert JJ, Joubert M, Nal L, Fieschi M. A computational Model of Information Retrieval with UMLS. Proc. 18th SCAMC. Ozbolt JG editor. JAMIA Symposium supplement. 1994: 167-71.
13. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf. Med. 1993; 32: 281-91.
14. Joubert M, Miton F, Robert JJ, Fieschi M. A Conceptual Graphs Modelling of UMLS Components. Proc. MEDINFO 95. Greenes R, Peterson H, Protti D, eds. North-Holland, 1995: 90-4.
15. Joubert M, Fieschi M, Robert JJ, Tafazzoli AG. Users Conceptual Views on Medical Information Databases. Int J Biomed Comput. 1994; 37: 93-104.